



# UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE  
United States Patent and Trademark Office  
Address: COMMISSIONER FOR PATENTS  
P.O. Box 1450  
Alexandria, Virginia 22313-1450  
[www.uspto.gov](http://www.uspto.gov)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/647,203	08/21/2003	Alexander Franz	24207-10274	1475
62296	7590	01/19/2007	EXAMINER	
GOOGLE / FENWICK SILICON VALLEY CENTER 801 CALIFORNIA ST. MOUNTAIN VIEW, CA 94041			SHAH, PARAS D	
			ART UNIT	PAPER NUMBER
			2112	
SHORTENED STATUTORY PERIOD OF RESPONSE	MAIL DATE	DELIVERY MODE		
3 MONTHS	01/19/2007	PAPER		

Please find below and/or attached an Office communication concerning this application or proceeding.

If NO period for reply is specified above, the maximum statutory period will apply and will expire 6 MONTHS from the mailing date of this communication.

<b>Office Action Summary</b>	Application No.	Applicant(s)
	10/647,203	FRANZ ET AL.
	Examiner	Art Unit
	Paras Shah	2112

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --  
**Period for Reply**

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

#### Status

- 1) Responsive to communication(s) filed on 21 August 2003.
- 2a) This action is FINAL.                    2b) This action is non-final.
- 3) Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

#### Disposition of Claims

- 4) Claim(s) 1-36 is/are pending in the application.
- 4a) Of the above claim(s) \_\_\_\_\_ is/are withdrawn from consideration.
- 5) Claim(s) \_\_\_\_\_ is/are allowed.
- 6) Claim(s) 1-4,6-9,11-18,20-24,26-29,31-36 is/are rejected.
- 7) Claim(s) 5,10,19,25 and 30 is/are objected to.
- 8) Claim(s) \_\_\_\_\_ are subject to restriction and/or election requirement.

#### Application Papers

- 9) The specification is objected to by the Examiner.
- 10) The drawing(s) filed on 8/25/2003 is/are: a) accepted or b) objected to by the Examiner.  
Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).  
Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).
- 11) The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

#### Priority under 35 U.S.C. § 119

- 12) Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
  - a) All    b) Some \* c) None of:
    1. Certified copies of the priority documents have been received.
    2. Certified copies of the priority documents have been received in Application No. \_\_\_\_\_.
    3. Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

\* See the attached detailed Office action for a list of the certified copies not received.

#### Attachment(s)

1) <input checked="" type="checkbox"/> Notice of References Cited (PTO-892)	4) <input type="checkbox"/> Interview Summary (PTO-413)
2) <input type="checkbox"/> Notice of Draftsperson's Patent Drawing Review (PTO-948)	Paper No(s)/Mail Date. _____.
3) <input checked="" type="checkbox"/> Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date <u>11/18/2003</u> .	5) <input type="checkbox"/> Notice of Informal Patent Application
	6) <input type="checkbox"/> Other: _____.

## **DETAILED ACTION**

1. This Office Action is in response to the Application filed on 08/25/2003.

### ***Specification***

2. The disclosure is objected to because of the following informalities: "L(H<sub>c</sub>)" on page 12 line 22 should be L(H<sub>i</sub>).

Appropriate correction is required.

### ***Claim Rejections - 35 USC § 112***

3. The following is a quotation of the second paragraph of 35 U.S.C. 112:

The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

4. Claim 13 recites the limitation "the limit" in line 10. There is insufficient antecedent basis for this limitation in the claim.

5. Claim 14 recites the limitation "the upper limit" in line 3. There is insufficient antecedent basis for this limitation in the claim.

### ***Claim Rejections - 35 USC § 103***

6. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

(a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

7. The factual inquiries set forth in *Graham v. John Deere Co.*, 383 U.S. 1, 148

USPQ 459 (1966), that are applied for establishing a background for determining

obviousness under 35 U.S.C. 103(a) are summarized as follows:

1. Determining the scope and contents of the prior art.
2. Ascertaining the differences between the prior art and the claims at issue.
3. Resolving the level of ordinary skill in the pertinent art.
4. Considering objective evidence present in the application indicating obviousness or nonobviousness.

8. Claims 1, 3, 6, 8, 11-13, 20-24, and 31-36 are rejected under 35 U.S.C. 103(a)

as being unpatentable over Su *et al.* (In *Proceedings of the 32nd Annual Meeting on Association For Computational Linguistics* 1994).

As to claims 1, 6, and 12, Su *et al.* discloses a system comprising of tokens (see page 244, Table 1) from a text corpus (see page 243, left column, 2<sup>nd</sup> paragraph, line 6). Su *et al.* further discloses compound finder iteratively finding compounds (page 244, left column, 1st paragraph, line 10) (e.g. It should be noted that windowing the corpus in sizes of 2 and 3 over the text corpus can be interpreted as a form of iteration when finding compounds of these various lengths) evaluating a frequency of occurrence (n-gram counter) (see page 244, left column, 1st paragraph, lines 3-4) for one or more n-grams (see page 243, left column, 3rd paragraph, lines 1-5). Also, Su *et al.* discloses a compound finder including an n-gram counter (see page 244, left column, 1<sup>st</sup> paragraph, lines 3-4) and a likelihood evaluator (see page 243, right column, line 8), which adds the compound words having a high likelihood to the vocabulary (see page 245, right column, 2<sup>nd</sup> paragraph, line 7). However, Su *et al.* does not specifically disclose a vocabulary comprising the tokens. It would have been obvious to one of ordinary skilled

in the art to have included vocabulary storage to store the tokens from a text corpus as shown by the reference. The motivation to have included such a unit involves the reference disclosing a token list from the text corpus of individual words. Thus, the token list must be stored in order to perform the compound search.

As to claims 3 and 8, Su *et al.* discloses a system where only some of the n-grams that have a high likelihood are added as compounds to the vocabulary (see page 245, right column, 2<sup>nd</sup> paragraph, line 6-8) (e.g. It should be noted that the selection of those compounds, which have a high likelihood will be chosen if the value is greater than 0, otherwise it will not be included).

As to claim 11, Su *et al.* does not specifically disclose the use of a computer for compound extraction. Su *et al.* does mention simulation for compound extraction (see page 245, right column, 2<sup>nd</sup> paragraph). Hence, it is obvious to one of ordinary skilled in the art to have used a computer to execute the simulation from code. The motivation to include a computer-storage medium is for use in machine translation (see page 243, left column, 1<sup>st</sup> paragraph, line 27).

As to claims 13, 24, and 36 Su *et al.* discloses a system for identifying compounds through iterative analysis comprising: the number of tokens per compound (see page 243, left column, 2<sup>nd</sup> paragraph, line 3 and line 10) (e.g. A limit is pre-specified by the reference); a compound finder evaluating compounds in a text corpus comprising: n-gram counter (see page 244, left column, 1<sup>st</sup> paragraph, lines 3-4) for determining number of occurrences of one or more n-grams (e.g. The maximum number of tokens depends on the iteration value or step); a likelihood evaluator (see

page 243, right column, line 8), which determines a measure of association between tokens (see page 243, right column, lines 20-23) and , which adds the compound words having a high likelihood to the vocabulary (see page 245, right column, 2<sup>nd</sup> paragraph, line 7). Further, the adjustment of the limit can also be interpreted as the change in the n value of an n-gram. Thus, the change of limit from n=2 to n=3, will change the number of tokens per compound (page 243, left column, 2<sup>nd</sup> paragraph, lines 9-10). However, Su *et al.* does not specifically disclose the use of a stored limit of the number of tokens per compound and the use of a vocabulary. It would have been obvious to one of ordinary skilled in the art to have included a predetermined limit on the number of token per compound and the use of a vocabulary. The motivation to modify the compound extraction by Su *et al.* by the inclusion of a stored limit is to acquire the compounds of interest to the user (see page 243; 2<sup>nd</sup> paragraph, line 6) (e.g. The reference uses n-grams of n=2, and n=3). The motivation to have included such a vocabulary involves the reference disclosing a token list from the text corpus of individual words. Thus, the token list must be stored in order to perform the compound search.

As to claims 20-21 and 31-32, Su *et al.* discloses a system where token are extracted from a text corpus (see page 243, left column, 2<sup>nd</sup> paragraph, lines 6-9) through morphological analysis (e.g. It should be noted that morphological analysis and parsing is similar). Further, Su *et al.* does not specifically disclose a vocabulary being constructed from the words obtained from morphological analysis. However, it would be obvious to one of ordinary skilled in the art to include the parsed words in a dictionary or

vocabulary for comparison (see page 246, left column, 2<sup>nd</sup> paragraph (Concluding Remarks), lines 5-10).

As to claims 22 and 33, Su *et al.* discloses where the number of occurrences of one or more n-grams within the text corpus for unique n-grams (see page 243, left column, 1<sup>st</sup> paragraph, line 3 and lines 7-9) (e.g. It should be noted that the use of the relative frequency is a measure for compound extraction and can thus be interpreted as a filtering means when the compound filtering is done) (see page 243, left column, 1<sup>st</sup> paragraph, lines 1-5).

As to claims 23 and 34, Su *et al.* discloses a system where the text corpus comprises of documents (see abstract). Su *et al.* does not specifically disclose the documents being a web page, new message, and text. However, Su *et al.* does indicate this can be used with machine translation (see page 246, left column (Concluding Remarks), line 1). It would have been obvious tone of ordinary skilled in the art to have included the mentioned documents. It should be further noted that a web page could consist of a news message, which contains text. Further, a machine translation of a website, which is a news page of another language can satisfy the incorporated reference.

As to claim 35, Su *et al.* does not specifically disclose the use of a computer for compound extraction. Su *et al.* does mention simulation for compound extraction (see page 245, right column, 2<sup>nd</sup> paragraph). Hence, it is obvious to one of ordinary skilled in the art to have used a computer to execute the simulation from code. The motivation to

include a computer-storage medium is for use in machine translation (see page 243, left column, 1<sup>st</sup> paragraph, line 27).

9. Claims 2, 7, 15 and 26 are rejected under 35 U.S.C. 103(a) as being unpatentable over Su *et al.* as applied to claims 1, 6, 13, and 24 above, and further in view of Takashi.

As to claims 2, 7, 15 and 26, Su *et al.* discloses the finding of compounds in a text corpus. However, Su *et al.* does not specifically disclose the use of an iterator used to count backwards from a set limit. Takashi discloses a similar type of iteration, where the n-gram is counted forward to a maximum value (see Page 2, [0006], in English translation) rather than backward. It would have been obvious to one of ordinary skill in the art to have modified the system by Su *et al.* with a backward counting mechanism as that by Takashi. This forward mechanism by Takashi could be changed to a backward iteration from a maximum (e.g. In the reference denoted as Nmax) since the same results would be evident due to the forming of the same word grouping pairs for (n=1, 2,3, where 3 is Nmax) (e.g. of word San Diego Zoo, Forward iteration yields: San, San Diego, San Diego Zoo; Backward iteration yields: San Diego Zoo, San Diego, and San) and a probability that is assigned by the likelihood ratio as evident by one of ordinary skill.

10. Claim 4, 9, 16-18 and 27-29 is rejected under 35 U.S.C. 103(a) as being unpatentable over Su *et al.* as applied to claims 1, 6, 13, and 24 above, and further in view of Manning (The MIT Press 1999).

As to claims 4, 9, 16-17, and 27-28 Su *et al.* discloses a system, where the likelihood ratio  $\lambda$  is computed by:  $\lambda = (P(x_c|M_c) * P(M_c)) / (P(x_{nc}|M_{nc}) * P(M_{nc}))$  (page 243, right column, line 9 (equation)) (e.g. It should be noted that the reference uses a different notation, but the same result and definitions are used, where the numerator is the n-gram produced by a compound result and the denominator is the result produced by a non-compound result. The formula can be changed to account for various distributions (Gaussian, Binomial). However, Su *et al.* does not specifically disclose the likelihood ration given by  $\lambda = L(H_i)/L(H_c)$ . Manning shows the use of the likelihood ratio (see equation 5.10)(e.g. The equation is given in log form. The logs can be omitted to obtain the desired formula. The numerator is the independent hypothesis and the denominator is the dependence hypothesis.) It would have been obvious to one of ordinary skilled in the art to have modified the formula by Su *et al.* with the formula presented by Manning. The motivation to modify the former is for collocation discovery (see Manning, page 172, sect. 5.3.4, 3<sup>rd</sup> paragraph, lines 1-4).

As to claims 18 and 29, Su *et al.* discloses a system for identifying compounds through measure of association. However, Su *et al.* does not specifically disclose the representation of the independence and collocation hypothesis. Manning does disclose the explanations of these two types of hypothesis (see page 172, sect. 5.3.4, bullet items) (e.g. It should be noted that the independence hypothesis is given by hypothesis 1 and the dependence or collocation hypothesis by hypothesis 2. The  $w_2$  and  $w_1$  can be interpreted as the tokens since the reference deals with a text corpus). It would have been obvious to one of ordinary skilled in the art to have included the

formulation of the hypothesis to that presented by Su *et al.* The motivation to modify the former is for collocation discovery (see Manning, page 172, sect. 5.3.4, 3<sup>rd</sup> paragraph, lines 1-4). Further, the use of the formula presented by Manning would require an explanation of frequency for each type of hypothesis in order to find the likelihood ratio (definition of likelihood ratio).

### ***Allowable Subject Matter***

11. Claims 5,10, 19, 25, and 30 are objected to as being dependent upon a rejected base claim, but would be allowable if rewritten in independent form including all of the limitations of the base claim and any intervening claims.
12. Claim 14 would be allowable if rewritten to overcome the rejection(s) under 35 U.S.C. 112, 2nd paragraph, set forth in this Office action and to include all of the limitations of the base claim and any intervening claims.
13. The following is a statement of reasons for the indication of allowable subject matter: none of the prior art references alone or in combination teaches or fairly suggests the limitations where “a limiter identifying a number of n-grams up to the upper limit based on number of occurrences” as seen in claims 14 and 25. Also, the limitations of “dividing the *n*-gram into *n*-1 pairings of segments... selecting the maximum likelihood of collocation of the pairings as L(H<sub>c</sub>)” as seen in claims 5 and 10. Further, the limitations “L(H<sub>i</sub>) is computed ... in accordance with the formula:

$$\arg \max_{L(H_i)} \frac{L(t_1, t_2 \text{ form compound})}{L(n - \text{gram does not form compound})} \text{ as seen in claims 19 and 30.}$$

***Conclusion***

14. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure.

The US 6,349,282 is cited to teach a compound word recognizer and a compound word detector using n-grams, respectively.

The NPL documents by Venkataraman and Gao *et al.* are cited to teach a method for extracting word sequences using n-grams and maximum likelihood principles.

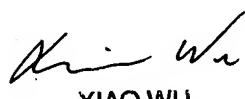
Any inquiry concerning this communication or earlier communications from the examiner should be directed to Paras Shah whose telephone number is (571)270-1650. The examiner can normally be reached on MON.-FRI. 7:30a.m.-5:00p.m. EST.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Xiao Wu can be reached on (571)272-7761. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see <http://pair-direct.uspto.gov>. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

P.S.

12/18/2006



XIAO WU  
SUPERVISORY PATENT EXAMINER